Human Voice Analysis and Virtual Teacher for Speech Therapy

Man-Ching Yuen[1]*, Chi-Wai Yung[1], Linjing Zhang[1], Jiaer Song[1], Xingzi Li[1] and Yinlin Li [1]

[1] Department of Applied Data Science, Hong Kong Shue Yan University, Hong Kong, China
mcyuen@hksyu.edu

Abstract. Based on the literature review, researchers reported that at most, 24.6% of young children in the world were estimated to have speech delay or speech sound disorder (SSD). Once children with SSD are identified, speech-language pathologists (SLPs) select initial therapy programs for children with regular review and adjustments on therapy. The success of therapy highly relies on the effectiveness of long-term home training. In this project, we carry out human voice analysis and design and implement a virtual teacher for home training in speech therapy. For the first part of this project, we conduct sound analysis research to see if children's Cantonese pronunciation is correct. Once the children's voices are captured, human voices can be automatically transferred for waveform analysis, allowing a large number of tasks to be completed quickly. The created waveform is compared to the standard waveform. If the majority of the waveform is inconsistent, it suggests that the pronunciation of children is not standard. As a result, it points out children's pronunciation problems and generates feedback quickly. Through the waveform diagram, our system can accurately process and analyze the sound, as well as eliminate the inaccuracy caused by varied timbres of children, making the analysis more accurate and effective. For the second part of this project, we implement a virtual teacher by using Blender and Audio2Face technology. Blender technology is often useful in areas such as live streaming and business, but it also has great potential in education. Therefore, it provides a more convenient way to conduct speech imitation and language learning for implementation of a virtual teacher. It can achieve low-cost popularization, timely correction of children's pronunciation problems.

Keywords: Speech therapy, Human voice analysis, Waveform analysis.

1 Introduction

1.1 Background

The diagnostic category of developmental speech sound disorders (SSD) poses a clinical problem due to its size, heterogeneous symptomatology, limited research base and poor long-term outcomes. SSD is the most prevalent of childhood communication difficulties, constituting more than 70 % of pediatric speech- language pathology case

A. C. Chan et al. (eds.), *Proceedings of the Positive Technology International Conference 2023 Positive Technology: Possible Synergies between Emerging Technologies and Positive Psychology (PT 2023)*, Atlantis Highlights in Social Sciences, Education and Humanities 19,
https://doi.org/10.2991/978-94-6463-378-8_7

loads. One UK study reported an incidence (referral) level of 6.4 % of the total child population aged 2–16 years. Once children with SSD are identified, speech-language pathologists (SLPs) select initial therapy programs for children with regular review and adjustments on therapy. The success of therapy highly relies on the effectiveness of long-term home training (Dodd, 2014).

In the SSD therapy process it had five principles (Liang, n.d.). The first is to intervene early, because infancy is the key period for a child's structural development of the nervous system. Second is to set a proper intervention target (Liu, 2022). In the SSD intervention, it will focus on different aspects, like children's basic social technology and language ability. In our project we focus on children phonetic therapy. The third is we need to consider the children's various characteristics. For example, the intervening environment and teaching method influence the children. The fourth is the practicability of the intervening method. The method should consider the things that children learning can use in daily life. The final one is repeat stimulation. With the help of repeat stimulation, children can find the right regulation in the process. It is also the most efficient way to improve language ability (Zhang, 2018).

In this project, we want to give high-risk SSD Cantonese children an early intervention in a convenient and efficient way. So, we will divide it into two parts: face and voice. For the voice analysis, we will compare the accurate pronunciation with the children's and compare whether their pronunciation is wrong. For the face part, we will use the 3D model to create a virtual teacher to guide and teach children.

1.2 Our Contribution

The major contributions of this work are as follows:

1) In order to effectively intervene in this symptom, we employ the contrast of the waveform rather than only an audio comparison with the norm to properly determine which syllable of children's pronunciation is out of the norm and directly correct it.
2) Compared with other usual 3D models, our 3D model is usually used in the medical field to teach the SSD children.
3) Compared with other animation virtual teachers, their mouth shapes can only be opened and closed simply, which cannot show the complexity of human mouth shapes, so we combine cartoon faces with slightly more complex mouths to show Create our own unique creations.
4) We mainly face high-risk SSD Cantonese children for an early intervention.

2 Related Work

2.1 Form of Data

With the definition of SSDs referring to difficulties in one or more aspects of perception, phonological representation, and motor production of speech sounds and segments (American Speech-Language-Hearing Association, n.d.)[1], the majority of the reviewed studies focus directly on the audial aspect of the speech sounds produced by experiment participants or targets in the form of audio files or audio in video files (Sztahó et al., 2018). However, there also are studies that collect data other than the audial dimension. In a study of classification of speech movements (Wang et al., 2016), data of the positions and movements of the tongue and lips were collected and analyzed instead of audial data using electromagnetic sensors. In a study of subclassification of SSDs (Vick et al., 2014), apart from audio data, articulatory data were also collected by recording the vertical movements of the upper lip, lower lip, and jaw by an infrared camera, light source, and circular reflective markers attached to the participants' face. In a wireless tongue tracking system (Sebkhi et al., 2021), data on the 3D position of the tongue were collected and analyzed using magnetic tracers.

2.2 Machine Learning in SSD Treatments

After a patient is diagnosed with a type of SSD, treatments are performed to help the patient. While different SLPs may adopt different treatment methods based on the situation of individual patients, machine learning techniques can be adopted to reduce the workload of SLPs or increase the efficiency of treatments.

2.3 Automatic Speech Recognition (ASR)

In Computer-Aided Pronunciation Training (CAPT) and Computer-Aided Speech and Language Therapy (CASLT), ASR systems can automatically measure the correctness in pronunciation and speech, providing real-time feedback to patients during training, and information to SLPs on the treatment progress and decisions of follow-up treatment methods. As described by (Saz et al., 2009), there are three steps in speech and language therapy: phonation acquisition, by which patients learn the pronunciations of individual phonemes; articulation acquisition, by which patients learn how to pronounce separate words; and language understanding, by which patients understand the meaning of words and sentences and are able establish meaningful dialogs. In all three stages, ASR systems can be fine-tuned and deployed to verify the correctness of utterness and speech in the form of an assessment tool (Ballard et al., 2019; Sztahó et al., 2018), or in the form of tools and serious games for the purpose of speech and language training (Saz et al., 2009; Duval et al., 2018).

2.4 Assignment of Activities and Exercises

Apart from the treatment directly, machine learning techniques and algorithms can also be adopted in the pre-treatment process. As proposed by upon grouping patients of similar disorders or other measures together with machine learning algorithms like

[1] American Speech-Language-Hearing Association. (n.d.). *Speech sound disorders - articulation and phonology*. https://www.asha.org/practice-portal/clinical-topics/articulation-and-phonology

clustering, the system can automatically assign the same activities and exercises to other patients in the same group, reducing repetitive work for SLPs.

3 Our Work

3.1 Software We Used to Analyze Human Voice

Audacity is a cross-platform audio editing software for recording and editing audio that is free and open source. Our team uses software to record the pronunciation of children, which can automatically form a waveform diagram and can be exported as a wav file.

Scilab Open-Source computing software is open-source software developed by INRIA-France (French National Institute of Information and Automation). Scilab is a kind of scientific engineering calculation software with rich data types, which can easily realize various matrix operations and graphical displays, and can be applied to scientific calculation, mathematical modeling, signal processing, decision optimization, linear/nonlinear control and other aspects. Our team uses this software to analyze wav files into spectrograms.

3.2 Software We Used to Create a Virtual Teacher

Blender is an open-source, multi-platform, lightweight all-round 3D animation software, it can support all 3D model creation processes. In the past, Blender has been used in TV advertisements and produces high quality short videos (Flavell, 2011). Blender can bring realistic 3D effects and bring more details to mouth shape. It can support various plugins. In the project, it can show the mouth shapes details.

Audio2Face comes preloaded with "Digital Markers". This is a 3D character model that can be animated with your audio track, just select and upload your audio. The technology feeds audio into a pretrained deep neural network, which outputs a 3D vertex mesh of the generated character to create facial animation in real time. You can also edit the character's performance by editing various post-processing parameters. The results you see on this page are mostly Audio2Face raw output with little or no edited post-processing parameters. The idea of Audio2Face is to use a person's training data to convert voice to expression, and then pass the person's expression to other faces.

As an essential part of human machine interaction, facial animation has been applied to many applications, such as computer games, animated movies, man-machine conversation, virtual human agents, etc. (Tian et al., 2019). It is helpful for our portrait modeling. And we will use Blender and Audio2face to correct the child's mouth shape.

3.3 Human Voice Analysis Process

For sound processing and analysis, our team examined starting consonants, vowels, and their various pronunciations in Cantonese, and analyzed each pronunciation into the form of a waveform diagram, which established a "library." Then the waveform diagram is analyzed as a spectrum diagram, and a data set of the spectrum diagram is made. When the students begin the test, their voices are captured and analyzed to create a spectrogram that is compared to the library's standard spectrogram. Through this process, by comparison, we can determine whether children's pronunciation is proper. Children with pronunciation faults can be correctly guided using the "library" correction of standard sounds and oral guidance from a virtual teacher.

3.4 Database Establishment

To establish a database for Cantonese pronunciation, we first gathered, sorted, and categorized all the syllables. Each syllable was then recorded and edited into a wav file. Since each word is composed of several syllables, we grouped related syllables together to form a library of words. We then analyzed each wav file using Scilab and created a spectrogram library by analyzing each syllable into a fixed waveform. This allowed us to determine the pronunciation standard for each syllable by examining its wave peak and span on the X-axis. We compiled a standard library of syllables to determine whether youngsters are pronouncing words correctly. This database will be used to help individuals improve their Cantonese pronunciation and to assist virtual teachers in teaching correct pronunciation.

3.5 Test Process

The child's pronunciation is captured using audacity once our recognition system is operational, and the speech is then processed to create a waveform graph as shown in Fig. 1. The wav file is then imported and analyzed as a spectrogram (Fig. 2). We can accurately determine which words and which syllables are not standard for children by comparing the corresponding syllables of each word with our existing "standard library." We can then play the existing files in the "library" to correct the pronunciation of children with the mouth movements of the virtual teacher.

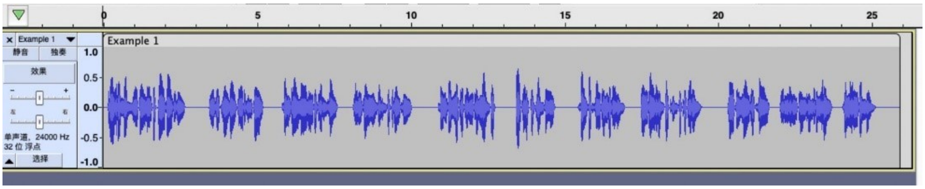


Fig. 1. Audacity Recording instance (Original waveform).

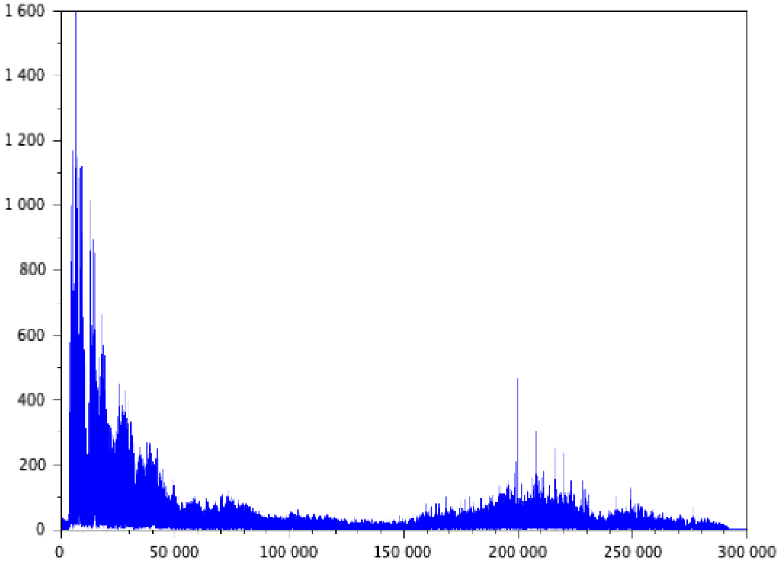
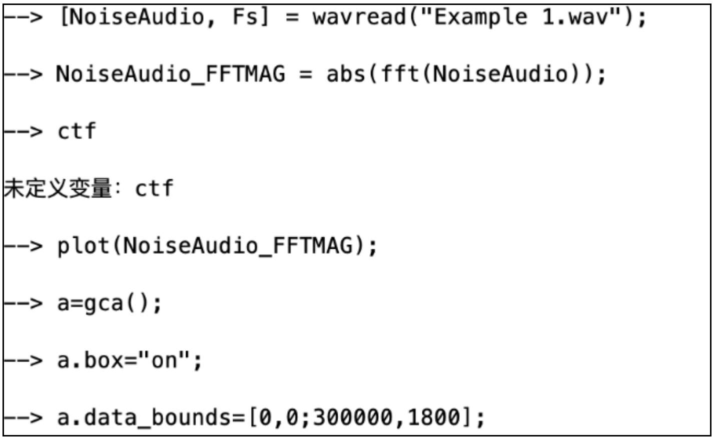


Fig. 2. Spectrogram after Scilab analysis.

3.6 Analyze Audio Files in Detail in Scilab

We use Scilab to read wav files and analyze them. Fig. 3 shows the codes for analyzing audio files in Scilab.



```
--> [NoiseAudio, Fs] = wavread("Example 1.wav");

--> NoiseAudio_FFTMAG = abs(fft(NoiseAudio));

--> ctf

未定义变量: ctf

--> plot(NoiseAudio_FFTMAG);

--> a=gca();

--> a.box="on";

--> a.data_bounds=[0,0;300000,1800];
```

Fig. 3. The codes for analyzing a wav file as a practical example of waveform diagram.

The process is further described as follows:

1. Read the wav file

```
[NoiseAudio, Fs] = wavread("C:\Users\User\Downloads\uk1.wav");
```

2. Convert audio to Spectrogram (Hz)

```
NoiseAudio_FFTMag = abs(fft(NoiseAudio));
```

3. Clear the last generated image

```
ctf
```

4. Chart the frequency spectrum (Hz)

```
plot(NoiseAudio_FFTMag);
```

5. Modify the x and y axes

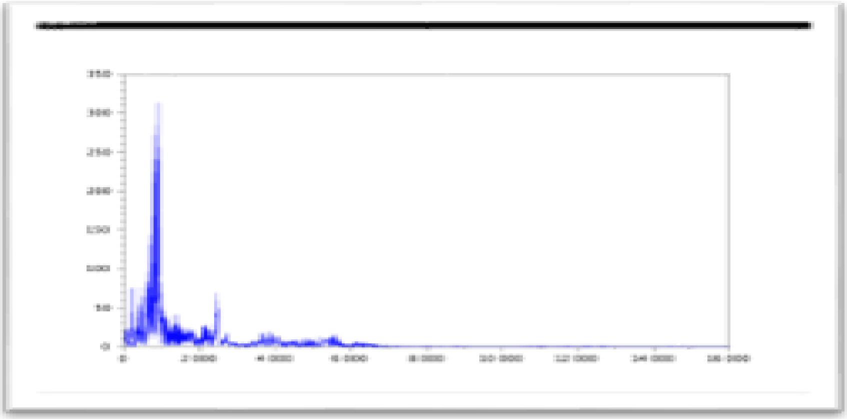
```
a=gca();
a.box="on";
```

6. Limit range [x_min,y_min:x_max,y_max]. The limit can be modified according to image range.

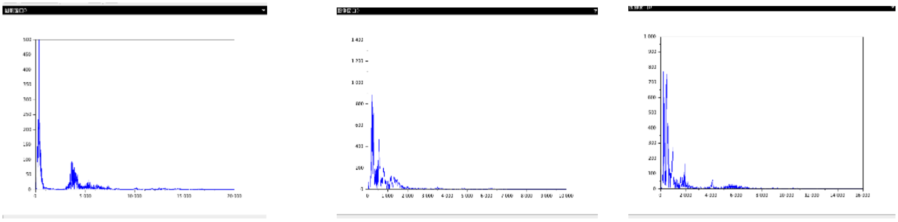
```
a.data_bounds = [0, 0, 5000 ，500];
```

3.7 Analyze Spectrograms of Correct and Incorrect Pronunciations

With the frequency value as the horizontal axis and the amplitude as the vertical axis, the amplitude of the above several sinusoidal signals is drawn on their corresponding frequencies, and the amplitude and frequency distribution diagram of the signal is made, which is the so-called spectrum diagram. Syllables are all three syllables, so the cycle is the same. The amplitude is the volume. Fig. 4. shows an example of what we tested. We know that "will" is made up of three syllables, will. The three pictures below are the spectrograms corresponding to the three syllables.



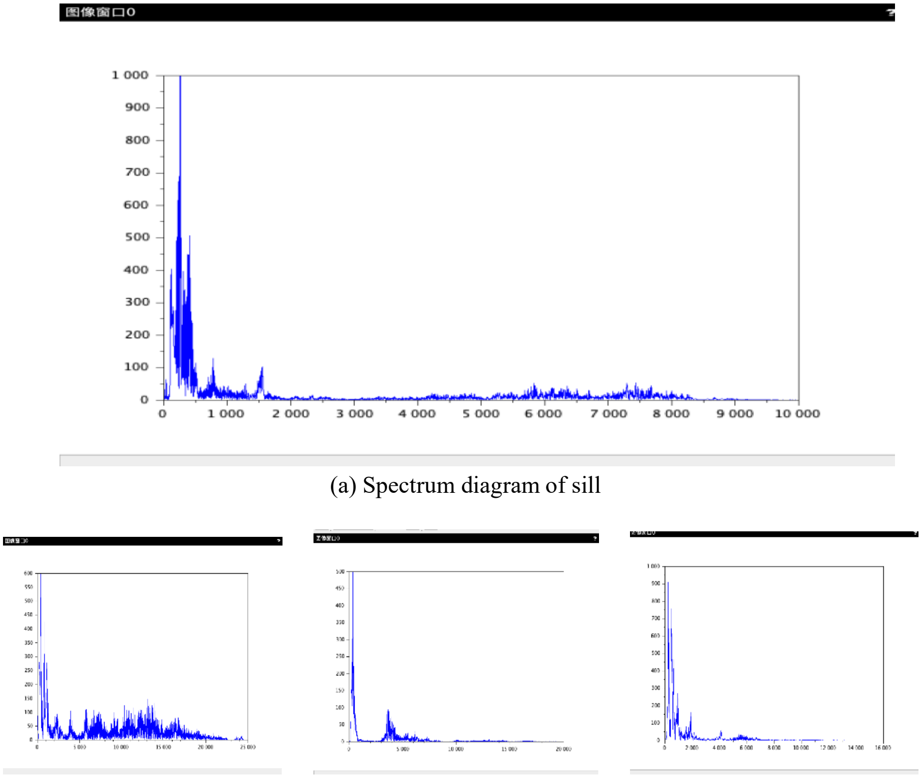
(a) Spectrum diagram of will



(b). Spectrum diagram of w (c). Spectrum diagram of i (d). Spectrum diagram of l

Fig. 4. Spectrogram after Scilab analysis of the word "will"

Next, we recorded a mispronunciation of sill when testing a toddler, who mispronounced a w as an s. Fig. 5 shows the spectrogram after Scilab analysis. The resulting spectrogram is very different from the will above. We can tell by contrast that his w sound is not pronounced correctly. It is possible to correct the pronunciation of young children.



(a) Spectrum diagram of sill

(b) Spectrum diagram of s (c) Spectrum diagram of i (d) Spectrum diagram of l

Fig. 5. Spectrogram after Scilab analysis when mispronunciation of the word "sill"

We can obviously see from the above spectrum diagram that the syllables of i and l are the same in will and sill, but the spectrum diagram of w and s is very different, which ultimately leads to a great difference in the spectrum diagram of will and sill. Therefore, we first form the spectrum diagram of will and sill to preliminarily judge whether the pronunciation is correct. Then, step by step, the spectrogram of each syllable is compared with the correct syllable to find the problem.

3.8 First Attempt at Creating a Virtual Teacher

In order to create a virtual teacher teaching SSD children, we tried two methods to create a virtual teacher. The first method is combining Photoshop (PS) and Live2D technology. We use Photoshop software to edit the picture and divide pictures into different parts. Put the edited picture into Live2D. After binding the character skeleton, we can use it to make an animation. But it cannot clearly show the change in mouth and teeth. And especially the mouth shape to achieve the relevant pronunciation, and the muscles in the relevant parts of the lips are difficult to display in 2D technology (Aneja & Li, 2019). Live2D mainly makes 2D animation (see Fig. 6). Compared with 2D technology, 3D will be a good improvement in the correction of pronunciation. And 3D technology can actually fine-tune the opening and closing of the relevant mouth and its muscles. So, we changed it to make 3D animation.

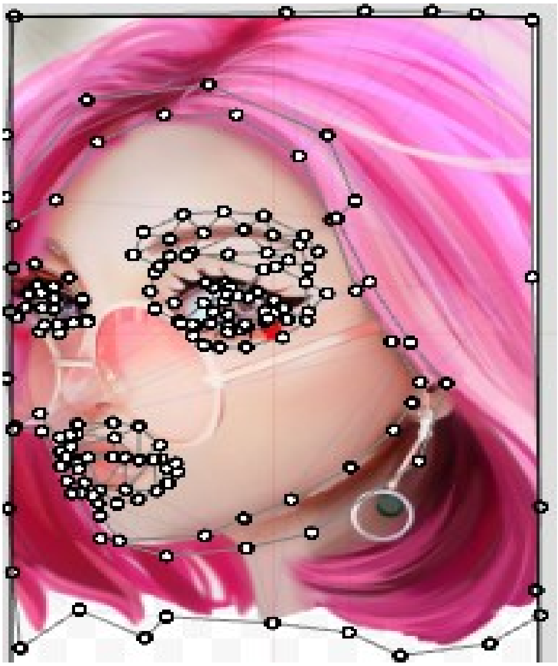


Fig. 6. Face skeleton in Live2D, picture comes from website.

3.9 Final Version about the Virtual Teacher

We use Blender to make 3D models. We have two different choices, one is to make a 3D human model, the other one is to make a 3D model cartoon character. Research shows that cartoon characteristics have stronger facial attractiveness than human characteristics (Chen et al., 2013). Above all, we choose to make a cartoon character as a virtual teacher. But due to technical reasons, we had to abandon the cartoon model and use the 3D human model instead. And create a human head and detail the human mouth, so that it can better show the sense of muscle lines when changing the mouth, so that children can better learn the details of the mouth when they pronounce it. In the 3D model we show the expected effect in the form of animation (see Fig. 7). After making a 3D model, we need to combine standard Cantonese pronunciation with mouth shapes. So, we found Audio2Face software which can combine voice and animation automatically.

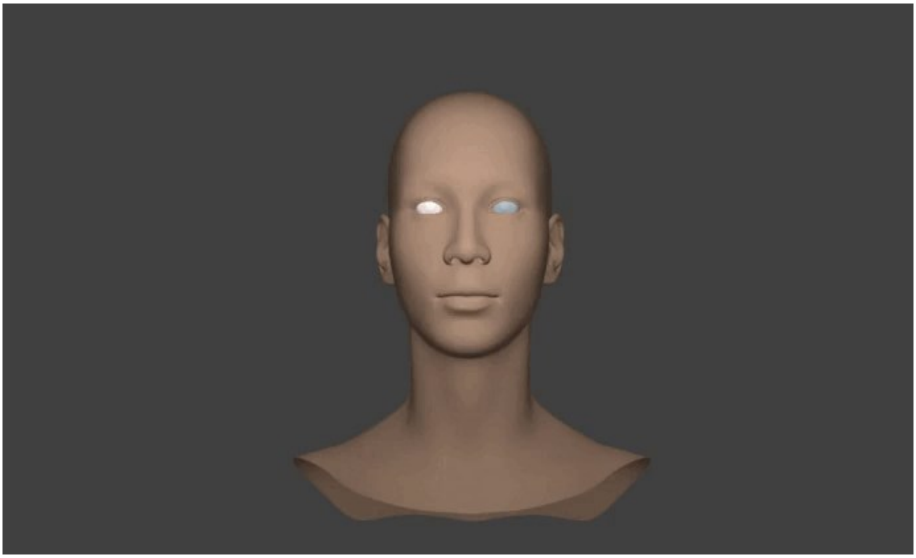


Fig. 7. Effect stimulation diagram (facial animation) in blender.

The idea of Audio2Face is to use a person's training data to convert voice to expression, and then pass the person's expression to other faces. Therefore, Audio2face will basically bring a lot of convenience to our work. We can put related models into it, which can realize the combination of audio and mouth shape, which has brought great help to our work. Audio2Face software can be used to combine the sound with the model made by Blender, so that the model of Blender can match the sound with the mouth shape, so as to have a virtual teacher's mouth shape animation. Our related models are shown in Fig. 8.

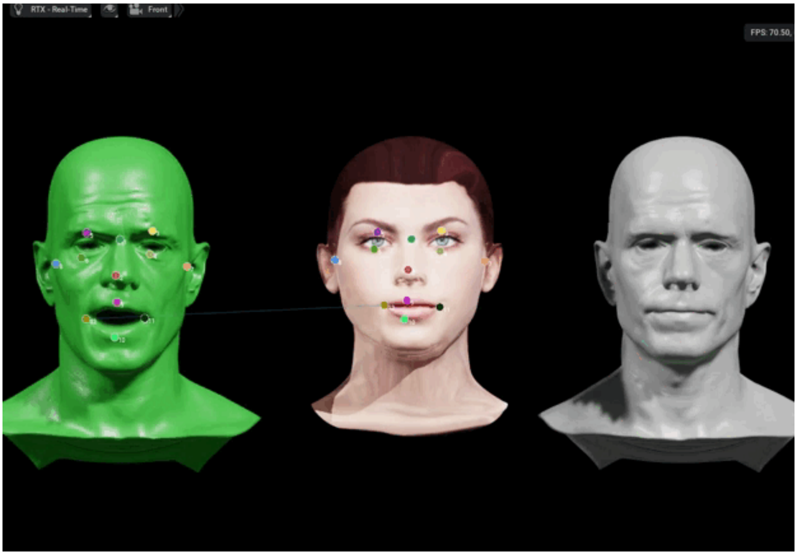


Fig. 8. Put the character model we made in Blender into Audio2face to match the mouth shape with the audio.

4 Conclusion and Future Work

In this project, we carried out human voice analysis, designed and implemented a virtual teacher for home training in speech therapy. However, the generated waveform map image at this stage was too small for us to clearly see the comparison between their waveform map and the correct waveform map, so we will amplify it using some programming techniques to get a sizable image. Besides, we cannot automatically determine the right waveform to compare with what the kids are saying, so the comparison takes a longer time. Last but not the least, our related models were not fully completed, so the opening and closing of the lips did not fully meet the expected level.

In our future plan, we will realize an online learning Cantonese platform for SSD children and achieve the mouth shape comparison.

Acknowledgements

This research was in part supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. UGC/FDS15/E01/21).

References

Aneja, D., & Li, W. (2019). Real-time lip sync for live 2d animation. https://doi.org/10.48550/arXiv.1910.08685

Ballard, K.J., Etter, N.M., Shen, S., Monroe, P., & Tan, C.T. (2019). Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia. *American journal of speech-language pathology*, *28*, 818-834, 10.1044/2018_AJSLP-MSC18-18-0109.

Chen, Y., Fang, H., Shen, N., Wu, M.-r., Zheng, L.-y., Wang, J.-m., & Lu, Y. (2013). Spatial relationship of facial features, effect of skin color brightness on the attractiveness of cartoon faces. *Psychological Development and Education*, *29*(6), 561-570.

Dodd, B. (2014). Differential diagnosis of pediatric speech sound disorder. *Current Developmental Disorders Reports, 1,* 189-196. https://doi.org/10.1007/s40474-014-0017-3

Flavell, L. (2011). *Beginning blender: Open source 3D modeling, animation, and game design*. Apress.

Liang, D. (n.d.). Education of children with language disorders [Review of education issues of children with language disorders]. *Chinese Journal of Social Sciences*, 2415.

Liu, X.-m. (2022). Clinical intervention of language delays and language disorders. *Chinese Journal of Child Health Care*, *30*(8), 813-817.

Saz, O., Yin, S., Lleida, E., Rose, R., Vaquero, C., & Rodríguez, W.R. (2009). Tools and technologies for computer-aided speech and language therapy. *Speech communication*, *51*, 948-967, 10.1016/j.specom.2009.04.006.Availableonline: https://dx.doi.org/10.1016/j.specom.2009.04.006.

Sebkhi, N., Santus, N., Bhavsar, A., Siahpoushan, S., & Inan, O.T. (2021). Evaluation of a wireless tongue tracking system on the identification of phoneme landmarks. *TBME, 68*, 1190-1197, 10.1109/TBME.2020.3023284. Available online: https://ieeexplore.ieee.org/document/9194333.

Sztahó, D., Kiss, G., & Vicsi, K. (2018). Computer based speech prosody teaching system. *Computer speech & language*, *50*, 126-140, 10.1016/j.csl.2017.12.010. Available online: https://dx.doi.org/10.1016/j.csl.2017.12.010.

Tian, G., Yuan, Y., & Liu, Y. (2019, July). Audio2Face: Generating speech/face animation from single audio with attention-based bidirectional LSTM networks. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, (pp. 366-371). IEEE. doi: 10.1109/ICMEW.2019.00069

Vick, J.C., Campbell, T.F., Shriberg, L.D., Green, J.R., Truemper, K., Rusiewicz, H.L., & Moore, C.A. (2014). Data-driven subclassification of speech sound disorders in preschool children. *Journal of speech, language, and hearing research, 57*, 2033-2050, 10.1044/2014_JSLHR-S-12-0193. Available online: https://www.ncbi.nlm.nih.gov/pubmed/25076005.

Wang, J., Samal, A., Rong, P., & Green, J. R. (2016). An optimal set of flesh points on tongue and lips for speech-movement classification. *Journal of speech, language, and hearing research, 59*, 15-26, 10.1044/2015_JSLHR-S-14-0112. Available online: http://eric.ed.gov/ERICWebPortal/detail?accno=EJ1092685.

Zhang, Y. (2018). Efficacy of guided education and sensory integration in children with language disorders. *China Minkang Science*, *30*(8), 85–87. https://doi.org/10 · 3969/j · issn · 1672－0369 · 2018 · 08 · 041